SEQUENTIAL VARIANCE ADAPTATION FOR REDUCING SIGNAL MISMATCHING

FIELD OF INVENTION

[0001] This invention relates to speech recognition and more particularly to mismatch between the distributions of acoustic models and noisy feature vectors.

BACKGROUND OF INVENTION:

[0002] In speech recognition, inevitably the recognizer has to deal with channel and background noise. The mismatch between the distributions of acoustic models (HMMs) and noisy feature vectors could cause degradation in performance of the recognizer. Model compensation is used to reduce such mismatch by modifying the acoustic models according to the certain amount of observations collected in the target environment.

[0003] Typically, batch parameter estimations are employed to update parameters after observation of all adaptation data which are not suitable to follow slow time varying environments. See L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE. 77(2): 257-285, February 1989. Also see C.J. Leggetter and P.C. Woodland, Speaker adaptation using linear regression, Technical Report F-INFENG/TR. 181,CUED, June 1994.

[0004] In recognizing speech signal in a noisy environment, the background noise causes the speech variance to shrink as noise intensity increases. See D. Mansour and B.H. Juang, A family of distortion measures based upon projection operation for robust speech recognition, IEEE Transactions on Acoustic, Speech and Signal Processing, ASSP-37(11):1659-1671, 1989.

[0005] Such statistic variation must be corrected in order to preserve recognition accuracy. Some methods adapt variance for speech recognition but they require an estimation of noise statistics to be provided. See M.J. Gales, PMC for Speech recognition in additive and convolutional noise, Technical Report TR-154, CUED/F-INFENG, December 1993.

SUMMARY OF INVENTION

[0006] In accordance with one embodiment of the present invention a method of updating covariance of a signal in a sequential manner includes the steps of scaling the covariance of the signals by a scaling factor; updating the scaling factor based on the signal to be recognized; updating the scaling matrix each time new data of the signal is available; and calculating a new scaling factor by adding a correction item to a previous scaling factor.

[0007] In accordance with an embodiment of the present invention sequential variance adaptation (SVA) adapts the covariances of the acoustic models online sequentially based on the sequential EM (Estimation Maximization) algorithm. The original covariances in the acoustic models are scaled by a scaling factor which is updated based on the new speech observations using stochastic approximations.

DESCRIPTION OF DRAWING

[0008] Figure 1 illustrates prior art speech recognition system.

[0009] Figure 2 illustrates the variance in a clean environment.

[0010] Figure 3 illustrates the variance for a noisy environment.

[0011] Figure 4 illustrates a speech recognition system according to one embodiment of the present invention.

DESCRIPTION OF PREFERRED EMBODIMENTS OF THE PRESENT INVENTION

[0012] A speech recognizer as illustrated in Figure 1 includes speech models 11 and speech recognition is achieved by comparing the incoming speech at a recognizer 13 to the speech models such as Hidden Markov Models (HMMs) models. This invention is about an improved model used for speech recognition. In the traditional model the distribution of the signal is modeled by a Gaussian distribution defined by μ and Σ where μ is the mean and Σ is the variance. The observed signal O_t is defined by observation N (μ, Σ).

[0013] Figure 2 illustrates the variance in a clean environment. Figure 3 illustrates the variance for a noisy environment. The variance is much narrower in a noisy environment. What is needed is to fix the variance to be more like the clean environment.

[0014] The mismatch between the distributions of acoustic models (HMMs) and feature vectors in speech recognition may cause performance degradation which could be improved by model compensation. Typically, batch parameter estimations are employed for model compensation where parameters are updated after observation of all adaptation data. Parameters updated this way are not suitable for follow slow parameter changes often encountered in speech recognition. Applicants' propose sequential variance adaptation (SVA) that adapts the covariances dynamically based on the sequential EM algorithm. The original covariances in acoustic models are adjusted by scaling matrices which are sequentially updated once new collection of data is available. SVA is able to obtain better estimation of time-varying model parameters to achieve good performance.

[0015] The following equation (1) is the performance index or Q function. The Q function is a function of θ which includes this bias.

$$Q_{K+1}^{(s)}(\Theta_k,\theta)=\sum_{r=1}^{K+1}Q_r(\Theta_k,\theta) \qquad (1)$$

where $Q_{k=1}^{(s)}$ denotes the EM auxiliary Q-function based on all the utterances from 1 to k+1, in which is the parameter set at utterance k and θ denotes a new parameter set. See A.P. Dempster, N. M. Laird, and D.B. Rubin "Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(1):1-38, 1977.

[0016] $Q_{k=1}^{(s)}$ can be written in a recursive way as:

$$Q_{k+1}^{(s)}(\Theta_k,\theta)=Q_k^{(s)}(\Theta_{k-1},\theta)+Q_{k+1}(\Theta_k,\theta), \qquad (2)$$

where $Q_{k=1}^{(s)}(\Theta_k,\theta)$ is the Q-function for the (k+1)th utterance. Based on stochastic approximation, sequential updating is

$$\theta_{k+1}=\theta_k-\left[\frac{\partial^2 Q_{k+1}^{(s)}(\Theta_k,\theta)}{\partial^2\theta}\right]_{\theta=\theta_k}^{-1}\left[\frac{\partial l_{k+1}(\Theta_k,\theta)}{\partial\theta}\right]_{\theta=\theta_k} \qquad (3)$$

[0017] Suppose the state observation power density functions (pdfs) are Gaussian mixtures with each Gaussian defined as equation 4.

$$b_{jm}(o_t)=N(o_t;\mu_{jm},\sum\nolimits_{jm}) \qquad (4)$$

$$=\frac{1}{(2\pi)^{\frac{n}{2}}\left|\sum\nolimits_{jm}\right|^{\frac{1}{2}}}e^{-\frac{1}{2}(o_t-\mu_{jm}-l_j)^T\sum_{jm}^{-1}(o_t-\mu_{jm})}$$

where the covariance matrix Σ_{jm} is assumed to be diagonal which implies the independence of each dimension of the feature vectors.

[0018] Since the components of feature vectors are assumed to be independent, the formulation on the sequential estimation algorithm is carried out using single variable for each dimension. The Gaussian pdf for the pth dimension in state j mixture m is

$$b_{jmp}(o_{t,p}) = N(o_{t,p};\mu_{jmp},\sigma^2_{jmp}) \qquad (5)$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{e^{\rho}{}_{p}\sigma^2_{jmp}}} e^{-\frac{(o_{t,p}-\mu_{jmp})^2}{2e^{\rho_p\sigma^2_{jmp}}}}$$

where the variance scaling factor e$^{\rho_p}$ takes an exponential form to guarantee the positiveness of the updated variances. The typical variance is $\sigma^2_{jmp.}$ We introduce e$^{\rho_p}$. ρ is a scalar number.

[0019] Also, to obtain reliable estimate, ρ's are tied for all phoneme HMMs for each dimension. But the derivation of ρ under alternate tying schemes is also straightforward. By computing the value of e$^{\rho_p}$ we can modulate the variance of any distribution. If this e$^{\rho_p}$ is larger you make the variance larger. We then try to optimally modify ρ so that we can find the best variance for the system.

[0020] Applying equation 3 with

$$Q_{k+1}(\Theta_k,\rho_p) = \sum_j\sum_m\sum_p\sum_{t=1}^{T^{k+1}}\gamma_{k+1,t}(j,m)\log b_{jmp}(o_{t,p}) \qquad (6)$$

$$= \sum_j\sum_m\sum_p\sum_{t=1}^{T^{k+1}}\gamma_{k+1,t}(j,m)\left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\rho_p - \frac{1}{2}\log\sigma^2_{jmp} - \frac{(o_{t,p}-\mu_{jmp})^2}{2e^{\rho_p\sigma^2_{jmp}}}\right]$$

where $\gamma_{k+1,t}(j,m) = P(\eta_t = j,\varepsilon_t = m \mid o_1^{T^{k+1}},\Theta_k)$ is the probability that the system stays at time t in state j mixture m given the observation sequence $o_1^{T^{k+1}}$, we get for second and first derivative

$$\frac{\partial Q_{k+1}(\Theta_k,\rho_p)}{\partial\rho_p} = \sum_j\sum_m\sum_{t=1}^{T^{k+1}}\gamma_{k+1,t}(j,m)\left[-\frac{1}{2} + \frac{(o_{t,p}-\mu_{jmp})^2}{2e^{\rho_p\sigma^2_{jmp}}}\right] \qquad (7)$$

$$\frac{\partial^2 Q_{k+1}(\Theta_k, \rho_p)}{\partial \rho_p^2} = -\sum_j \sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \frac{(o_{t,p} - \mu_{jmp})^2}{2e^{\rho_p \sigma_{jmp}^2}} \tag{8}$$

and the sequential updating equation is finding older ρ plus adjustment quantity as

$$\rho_p^{(k+1)} = \rho_p^{(k)} + \left[\sum_j \sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \frac{(o_{t,p} - \mu_{jmp})^2}{2e^{\rho_p \sigma_{jmp}^2}} \right]^{-1} \cdot \left[\sum_j \sum_m \sum_{t=1}^{T^{k+1}} \gamma_{k+1,t}(j,m) \left[-\frac{1}{2} + \frac{(o_{t,p} - \mu_{jmp})^2}{2e^{\rho_p \sigma_{jmp}^2}} \right] \right] \tag{9}$$

[0021] The above equation 9 states that the updated scaling factor is the current scaling factor plus a correction, which is a product of two factors.

[0022] After every utterance an update is done so that it is sequential. As illustrated in Figure 4 the steps according to the present invention are an utterance is recognized, the variance is adjusted using the utterance and then the model is updated. The updated model is used in the recognition of the next utterance and the variance is adjusted using the previously updated value plus the new adjustment quantity. The model is then updated.

[0023] The method of updating covariance of a signal in a sequential manner is disclosed wherein the covariance of the signal is scaled by a scaling factor. The scaling factor is updated based on the signal to be recognized. No additional data collection is necessary. The scaling factor is updated each time new data of the signal is available. The new scaling factor is calculated by adding a correction item to the old scaling factor. The scaling factor can be a matrix. The scaling matrix could be any matrix that ensures the scaled matrix a valid covariance. The new available data could be based on any length, in particular, it could be frames, utterances or every 10 minutes of a speech signal. The correction is the product of any sequences whose limit is zero, whose summation is infinity and whose square summation is not infinity and a summation of quantities weighted by a probability.